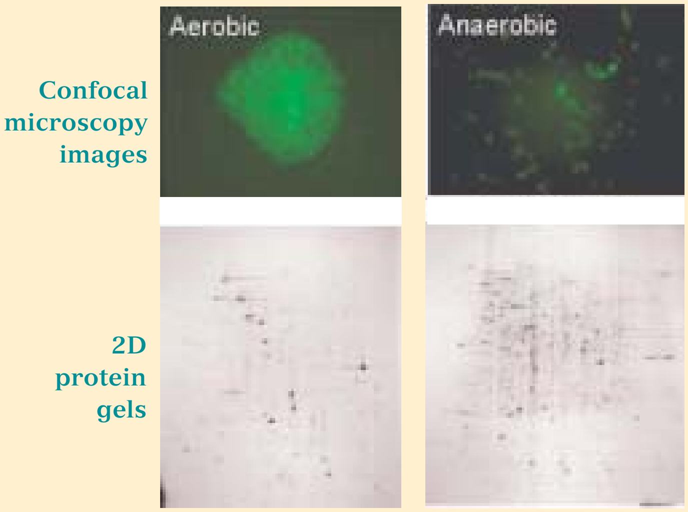
A Possible Application of Knowledge Gained from GTL Facilities

Shewanella oneidensis: *Offering New Strategies for Groundwater Bioremediation*



Proteome of aerobic and anaerobic *Shewanella* cultures

The ability of *S. oneidensis* to precipitate radionuclides (e.g., uranium and technetium) and metals (e.g., chromate), rendering them immobile in sediments, offers exciting new opportunities for developing new groundwater bioremediation strategies to clean up DOE legacy wastes. A GTL pilot project involving a consortium of scientists is now exploring how the microbe senses and responds to its environment. The team is cultivating the microbe under variable but carefully controlled conditions and measuring global (population) changes in gene expression and the resulting proteome. This work is enabled by the availability of the microbe's genome sequence, determined in the Office of Science's Microbial Genome Program.

Pictured is the growth of cells (top row) and protein complement (bottom row) present in the microbe under different environmental conditions. The pattern of cellular growth appears very different under aerobic (minibiofilm) and anaerobic (individual cells) conditions. These differences are reflected in the proteins revealed by 2D gels of the two cell cultures. An ability to quantitatively and comprehensively identify the cellular proteins will be critical for determining how *S. oneidensis* responds at the whole-system level.

Facility III: Characterization and Imaging of Molecular Machines

Cells are biological "factories" that perform and integrate thousands of discrete and highly specialized processes through the coordinated use of molecular "machines" composed of assemblies of proteins and other molecules.

Facility III will isolate, identify, and characterize from microbes thousands of molecular machines and develop the ability to image component proteins within complexes and to validate the presence of the complexes within cells.

Strategic Intent

Proteins seldom act in isolation; instead, they combine to form multiprotein complexes that function as "molecular machines." The genome and its associated sensing and regulatory networks control the creation and operation of the numerous molecular machines that make up a cell. These molecular machines are in turn organized into numerous tightly packed and highly interconnected physical structures. The dynamic interactions of proteins comprising many hundreds of these molecular machines are coordinated in time and space and are responsible for signaling, transport, motility, cell division, and virtually all other cell activities. Before we can progress toward a systematic understanding of cell function, we must discover which molecular machines can be produced by the cell under specific conditions and how they are positioned in the cell's structural architecture.

This facility's core role is to build on the data and reagents provided by Facility I and patterns in protein expression from Facility II to develop a detailed descriptive understanding of how proteins are organized into "molecular machines" and to locate the machines in the cell. The data from Facility III—together with dynamic snapshots of the complete, condition-dependent, expressed proteome to be generated by Facility II—will constitute the foundation on which Facility IV will develop a predictive, systems-level understanding of microbial cells and communities.

As important as protein complexes are in cellular function, our current knowledge of molecular machines is quite limited, partly because proteins most often have been studied individually and in isolation. Inherently difficult to study, many complexes are short-lived, unstable, or variable in their composition. In addition to identifying complexes, characterizing interactions among components will be critical to understanding their function. A first step toward this goal can be taken by determining the proteins comprising each complex and how the proteins interact to affect the machine's function.

Facility III will identify molecular machines and their components, characterize the interactions of the protein components of the complexes, and validate the occurence of these within the cell context. To accomplish this for the comprehensive set of molecular machines within the cell, the facility must develop automated analytical techniques for purifying, identifying, and characterizing multiprotein complexes—particularly, advanced imaging techniques. Integration, organization, and analysis of the data generated in this facility will require further development of principles, theory, and new computational tools for modeling and simulation of the structure and function of the complexes. Moreover, this facility will use the vast wealth of data on individual proteins being produced by structural genomics programs in other agencies including NIH and NSF.

Project Purpose and Justification

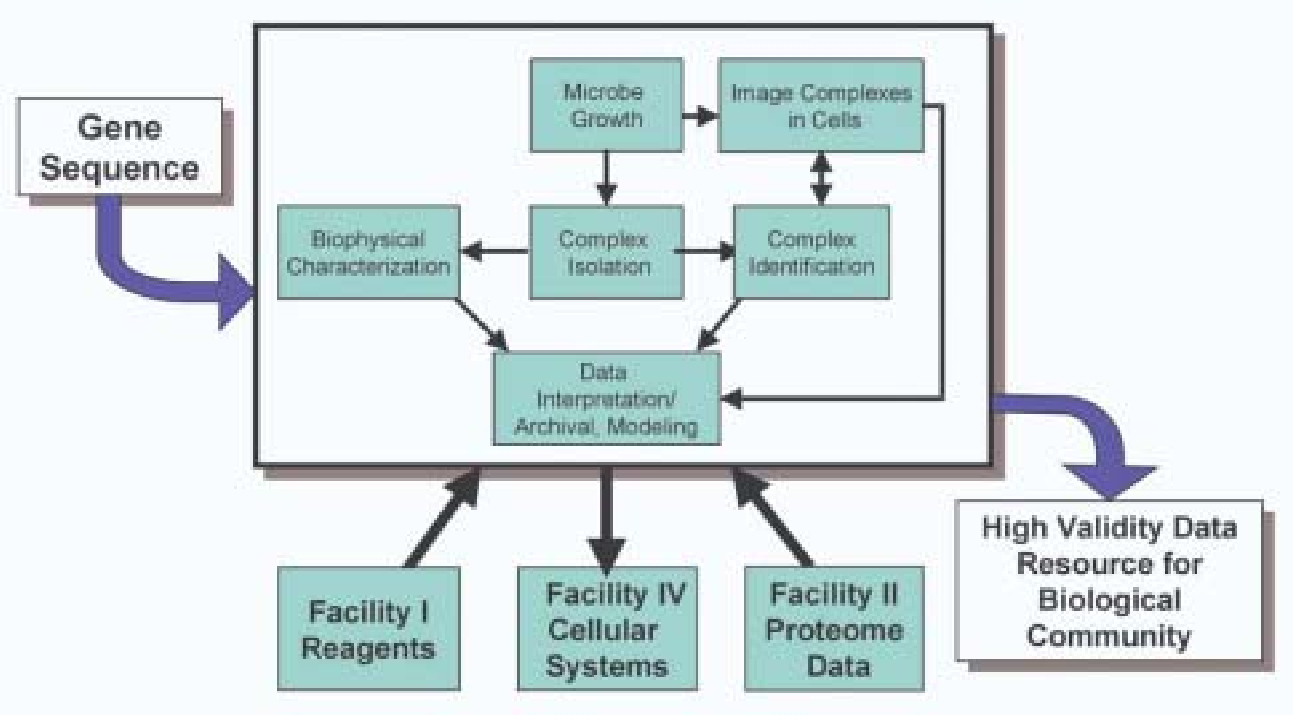
Molecular machines are highly dynamic, changing in composition, modification state, and subcellular location to carry out the vital functions of a cell. Responsible for a hierachy of molecular processes within and between cells, they dictate how a cell or organism interacts with its environment. A first step in determining how the network of cellular molecular processes works on a whole-systems basis is to completely understand individual molecular machines, how each machine is assembled in 3D, and how it is positioned in the cell with respect to other components of cellular architecture. Moreover, understanding how molecular machines operate at the molecular level will unlock the capability to control useful biochemical processes in a microbe and apply them to DOE mission needs.

This facility will focus on the isolation, identification, and characterization of thousands of molecular machines per year, providing revolutionary new information and capabilities to the biological community. Specifically, the facility will

- Discover and define the repertoire of protein complexes in a comprehensive manner; in this, it will rely critically on the information and reagents produced by Facilities I and II.
- Characterize the complexes as to their basic biophysical properties and also their inter-protein geometries.
- Develop molecular-level models to help interpret experimental data on protein interactions and modes of action of multiprotein machines.
- Develop theoretical principles for systematically describing the structure, aspects of function, assembly, and disassembly of multiprotein complexes.
- Validate the occurrence of complexes within the cells.

To reach the goal of the Genomes to Life program, vast numbers of data sets must be acquired from organisms maintained under a variety of well-defined conditions, then analyzed and made available to the biological community. Centralizing these analyses within a specialized facility analogous to today's genome-sequencing centers would allow assays to be conducted with higher efficiency, fidelity, and cost-effectiveness than could be accomplished in the laboratories of individual investigators.

Facility III: Characterization and Imaging of Molecular Machines


Gene Sequence
Microbe Growth
Image Complexes in Cells
Biophysical Characterization
Complex Isolation
Complex Identification
Data Interpretation/ Archival, Modeling
High Validity Data Resource for Biological Community
Facility I Reagents
Facility IV Cellular Systems
Facility II Proteome Data

Facility III will isolate, identify, and characterize from microbes thousands of molecular machines and develop the ability to image component proteins within complexes and to validate the presence of the complexes within cells. Facility III requires the proteins and reagents from Facility I for the isolation, quantitation, and imaging of protein complexes. Facility II will provide baseline proteomics data for optimizing conditions for production of machines to be analyzed in Facility III. The detailed structural and biophysical characterization of the complete repertoire of a cell's molecular machines, provided by Facility III, is critical to understanding and modeling cellular systems in Facility IV.

Project Description

Facility III will house state-of-the-art analytical instrumentation for the identification and characterization of molecular machines. The instrumentation would include electron, optical, and force microscopes; mass spectrometers; and other analytical tools. Laboratories also will be required for microbial cell growth, molecular biology, high throughput, automated sample preparation, gene expression, mass spectrometry–based protein complex analysis, imaging of protein complexes, biophysical characterization, and quality assurance. Integrated with these facilities will be computing resources for sample tracking; data acquisition, storage, and dissemination; algorithm development; and modeling.

For multiprotein machines with structurally characterized components, high-performance computing will play a very significant role in constructing structural models of machines and performing molecular dynamics simulations of protein-protein interactions in molecular machines. The next generation of massively parallel processors in the 40- to 100-teraflop range will allow simulations of sufficient size and fidelity to make important contributions to explaining the mechanisms of machine construction and function.

In summary, Facility III will

- Isolate complexes from cells using high-throughput techniques.
- Identify molecular components of the complexes.
- Determine basic biophysical properties of the complexes.
- Interpret, annotate, and archive data for use by the greater biological community.
- Develop models for the assembly and activity of complexes and verify this information with experimental data.

Molecular Machine Isolation

Perhaps the most challenging task in the analysis of molecular machines is the isolation of these complexes from the cell. Protein complexes often are held together by only weak interactions, making them fragile and difficult to isolate for analysis. Many such complexes are present only briefly or in very low amounts—sometimes just a few per cell. No adequate techniques now exist for the robust, high-throughput isolation of protein complexes. The development and automation of such techniques is therefore an essential early goal of a current GTL pilot project for this facility. Data and reagents to be produced by Facility I will be central to isolating the multiprotein complexes. In particular, Facility I reagents such as antibodies or clones, intended to produce "tagged" proteins in Facility III, will be used to purify complexes from cells by "pull-down" experiments. These methods, however, must be highly automated to meet the ultimate goals of comprehensively identifying the multiprotein machines in a cell. Automated techniques will be established for final purification (i.e., desalting, buffer exchange, and sample concentration), stabilization, storage, and proteolytic digestion of samples as required for analysis. Novel techniques for analyzing protein complexes in single microbes also will be developed.

An important component of this facility is a highly integrated LIMS that will track samples from cell cultivation through data archiving.

Mass Spectrometry

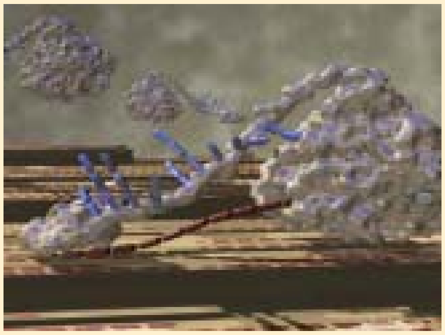
MS techniques provide the primary approach for analyzing the components of a molecular machine. This application, however, presents significant challenges. Spectrometers are required that add high-throughput operation to the current state-of-the-art standard in sensitivity, dynamic range, and resolving power. Mass spectra obtained from molecular complexes will produce data that can be evaluated on the basis of genomic-sequence data from the organism producing the complex and from proteomics data generated in Facility II. Typically, tens of thousands of spectra will be necessary to identify all protein components of a single multiprotein machine. Extensive computational analyses using genomic and proteomics data will be necessary to interpret these vast amounts of mass spectra from the complexes.

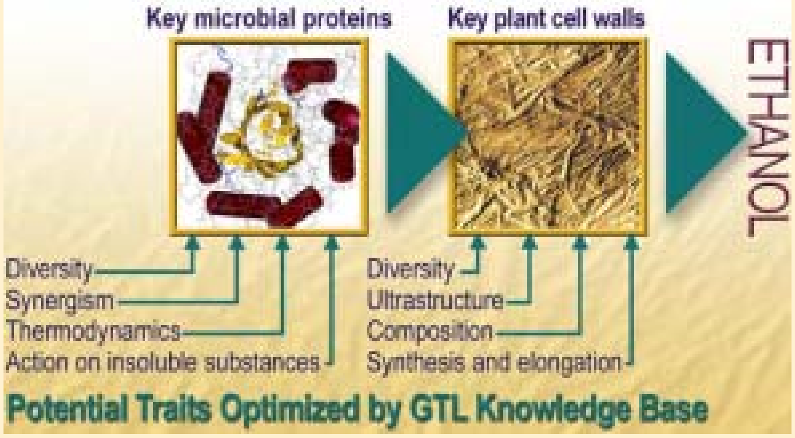
Biophysical Characterization

Generating isolated molecular complexes offers a unique but extremely challenging opportunity to characterize the complex with a host of biophysical techniques toward the ultimate goal of fully understanding the multiprotein machine's activity and mechanisms. Initially, a suite of well-established techniques will be employed to characterize the basic biophysical properties of an isolated complex. These techniques include surface plasmon resonance to investigate intermolecular interactions between the complex and weakly bound ligands; small-angle X-ray and neutron scattering to characterize overall structure and compactness of the complex; and wide-angle

A Possible Application of Knowledge Gained in GTL Facilities

Clean, Sustainable Energy





Enhanced plant qualities and microbial bioprocesses can be used to generate clean, sustainable energy. Microbial protein "machines" can break down the cellulose in plant cell walls for fermentation to ethanol. Today, the process is too inefficient for commercial production. Fundamental knowledge of gene regulation and protein machines gained in GTL can be applied to develop highly efficient methods to support large-scale ethanol production and displace a significant amount of fossil fuel use.

X-ray scattering to characterize secondary structure and folding. New technologies just being developed—such as single-molecule spectroscopy—are expected to allow more complete mechanistic understanding of multiprotein machines. Obtaining systematic experimental information about the dynamic behavior of the complexes (including their assembly and disassembly), combined with ongoing improvements in computational hardware and modeling methods, will allow accurate simulations of the activities of multiprotein machines at the heart of cellular function.

Imaging

Two different technologies will use (1) very high resolution to derive detailed 3D information about the complexes and (2) cell imaging to localize these complexes in individual cells.

Detailed 3D information about the structural organization of isolated molecular complexes will require many imaging technologies, including cryoelectron microscopy and a diversity of scanning and force microscopies. These tools will allow us to formulate the 3D structure of the complexes and provide hints as to how proteins interact.

An important application of imaging tools will be to verify the formation of complexes identified by MS and to map their location in the cell. Affinity reagents acquired from Facility I can be used to tag specific components of the complex and identify the location of complexes within the cell as well as the dynamics of assembly and disassembly. Electron tomography, X-ray microscopy, and live-cell imging using various light microscopy techniques will yield this data and will validate (in the cell) the protein interactions inferred from MS, biophysical, and 3D structure-imaging results. This information will provide additional insight into understanding the function of protein machines and will furnish valuable data for system-wide studies to be conducted in Facility IV.

Imaging methodologies produce vast amounts of raw data that must be analyzed computationally to produce interpretable images. Computer systems capable of acquiring and processing huge data streams must be assembled, and new algorithms must be developed to analyze and integrate data from multiple imaging modalities. Finally, strategies must be designed for archiving and distributing image data to the biological community.

Technology Development

As outlined previously, available methods for the isolation and MS-based analysis of protein complexes are not immediately adaptable to very high throughput operations. Thus, a key component of this facility will be the development of more robust and automated biological, analytical, and computational tools to improve sample throughput and information capture from these techniques.

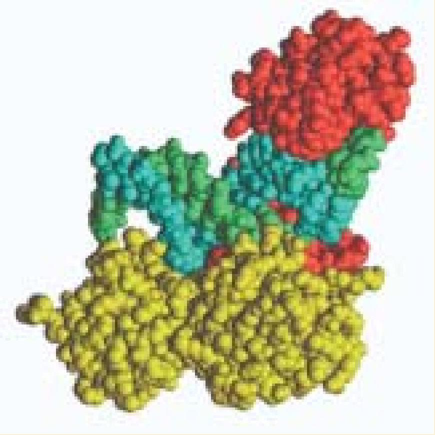
Impacts on Science and DOE Missions

Facility III will enable a fundamental understanding of the repertoire and properties of molecular machines present in cells. This is a prerequisite to determining how a microbe's network of molecular machines controls biochemical processes and therefore microbial life. It is the foundation on which we will achieve a fundamental scientific understanding of microbial life as well as the practical mastery needed to address DOE's mission goals in environmental remediation, energy production, and carbon cycling.

Probabilities for Success

- Analytical technologies have long been a strength of DOE national laboratories, and new technologies to analyze simple protein complexes are already in development in pilot projects funded by BER.
- DOE laboratories have a long history of excellence in high-performance computation and predictive simulations that can be applied to the study of molecular machines.
- Ongoing developments across many agencies and disciplines in robotics, automation, and data-management techniques will provide improved technologies. These technologies will be used to analyze the thousands of molecular machines in a microbe and identify biochemical pathways and gene regulatory networks that confer specialized capabilities possessed by microbes studied in GTL.
- Cryoelectron microscopy is a well-proven technique for analyzing the structure of molecular machines. Its development over the past 25 years has involved significant efforts at several national laboratories, including Brookhaven and Berkeley. High-end computing will enable integration of thousands of molecular images into progressively higher-resolution images of these complexes.
- New strategies using MS are being developed for isolating, stabilizing, and analyzing molecular complexes in one of GTL's initial projects. Oak Ridge National Laboratory is employing automated, high-throughput technologies and computational tools to capture these complexes for study. Partners are Pacific Northwest, Sandia, and Argonne national laboratories; University of Utah, and University of North Carolina.

Understanding Molecular Machines Requires Ultrascale Computing



Computationally simulating molecular machine activity—a critical prelude to understanding and using microbial capabilities—requires levels of computing power far beyond that available today. Simulating just the key steps occurring during the activity of the DNA-binding complex in *Pyrococcus woesei* (image at left) would require 40 teraflops of computing muscle—beyond the limit of what scientists can do now. The machine is made up of transcription-initiation factor II B (yellow), the TATA-box binding protein (red), and DNA (green and blue). Obtaining a complete simulation of the full activities of this or other complex molecular machines would require a capability of more than 100 and perhaps as many as 1000 teraflops, which are not available today (see sidebar, p. 33), but are on the planning horizon for OASCR.

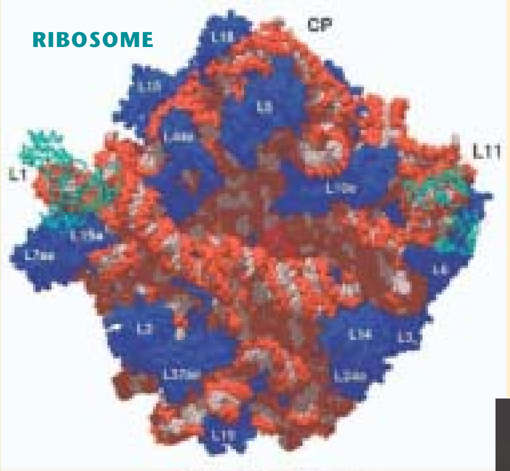
Tiny Molecular Machines Outstrip the Best Synthetic Chemists

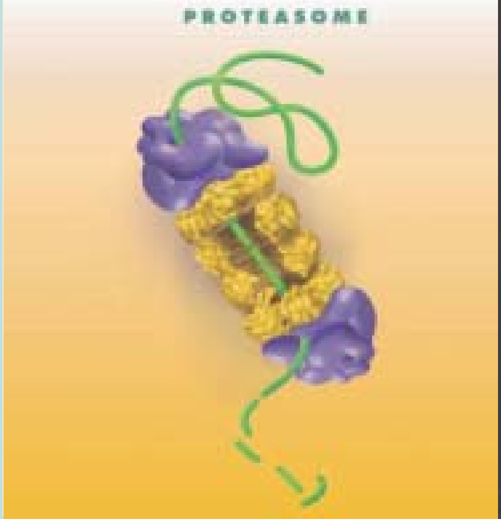
Molecular machines—assemblages of proteins and other chemical components—underlie the dynamic life of a cell. They work together in functional networks to carry out most life processes, including executing metabolic functions, mediating information flow within and among cells, and building cellular structures.

The ribosome (top image, below) is the molecular machine responsible for protein synthesis in cells. This remarkable molecular machine translates the sequence of bases on a messenger RNA molecule (a transcript of the DNA sequence) into a parallel sequence of amino acids that make up a protein chain. The ribosome carries out all the chemical reactions needed to perform these tasks.

On an atomic scale, the ribosome is huge. Made up of more than 50 proteins and 3 or 4 strands of RNA, it contains over 100,000 atoms. Understanding its function—determined by the ribosomal structure—was one of the hottest challenges in biology from the 1970s until a few years ago, when the first high-resolution structures began providing some insights into how the ribosome conducts its activities. This work required herculean efforts by hundreds of scientists in numerous laboratories. Data collected at the synchrotrons at Brookhaven, Berkeley, and Argonne national laboratories were used to calculate the high-resolution 3D structure of the ribosome's large and small subunits. The availability of genome data and current mass spectrometry technologies would now enable this result to be achieved at a fraction of the time and cost.

In addition to making proteins, cells also must have a way to destroy or recycle them. Sometimes this is important for maintaining a balance of proteins and amino acids as cells respond to external stimuli. At other times, cells need to degrade misfolded proteins to reuse their amino acids and also to keep the defective proteins from interfering with cellular activities. A cell must have a carefully regulated way of deciding which proteins to degrade, so it makes sense that all cells have special molecular machines—proteasomes—to degrade proteins. Made up of roughly 30 proteins, this protein complex also is very large. The structure of the central portion has been solved at high resolution by protein crystallography, and the whole structure has been imaged with cryoelectron microscopy (photo at left).





Facility IV: Analysis and Modeling of Cellular Systems

The final step in achieving a comprehensive understanding of living systems will require the ability to measure and predict dynamic events within individual cells.

Facility IV will combine advanced computational, analytical, and experimental capabilities for the integrated observation, measurement, and analysis of the spatial and temporal variations in the state of cellular systems—from individual microbial cells to complex communities and multicellular organisms.

Strategic Intent

The ultimate empirical test of our understanding of genomes will reside in our ability to model and simulate entire living systems, from individual microbial cells to complex microbial communities and, eventually, multicellular organisms. Understanding how individual cells within populations and in multi-organism communities interact and function as a unit to carry out complex processes is key to unlocking their vast potential for applications of importance to DOE.

The ability of our planet to sustain all life is completely dependent upon microbes. They are the foundation of the biosphere, controlling biogeochemical cycles and affecting the productivity of our soils, the quality of our freshwater supplies, and local and global climates. Microbes carry out sophisticated biochemical functions to degrade wastes and organic matter, cycle nutrients, and, as part of the photosynthetic process, to convert sunlight into energy and "fix" CO_2 from the atmosphere.

In nature, microbes often live in communities containing many different species. Yet, biologists traditionally have studied microbes one species at a time in nutrient-rich media, conditions typically very different from the organism's native habitat. To complicate matters further, microbes can exhibit substantial cell-to-cell phenotypic variation even in pure culture, necessitating the measurement of cell properties and functions at the level of the individual cell. Biologists are now constrained by the ability to make measurements on individual living cells and in microbial assemblies have acknowledged that instrument development and new facilities are critical needs in microbial science.

To make the "final ascent" to a systems-level understanding of life, new and innovative capabilities are needed for the comprehensive characterization of dynamic cellular systems at the level of the individual cell and in the context of their environment.

To this end, Facility IV will combine advanced computational, analytical, and experimental capabilities for the integrated observation, measurement, and analysis of the spatial and temporal variations in the state of microbial cells. To achieve a systems-level understanding, simulation and modeling must be tightly coupled with experiment to define and analyze the complex regulatory and metabolic networks in microbial cells and communities. Facility IV, building on knowledge and materials from Facilities I through III, will provide new analytical and computational tools and infrastructure that will enable unprecedented insights into the cellular state of microbes in populations and, ultimately, communities and multicellular organisms.

Facility IV will utilize advanced instrumentation and will be data and compute intensive, providing linked data sets on the dynamic function and behavior of single-cell and multiorganism assemblages and the capabilities for developing and evaluating them.

In summary, Facility IV will

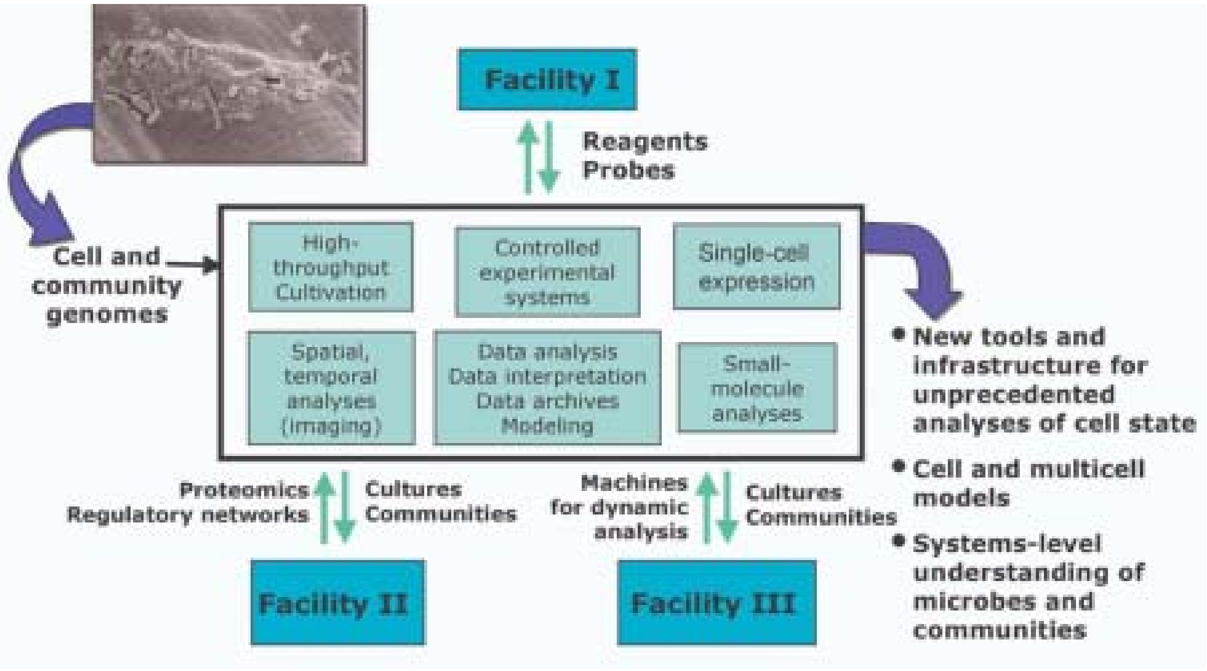
- Develop highly controlled systems for growing and maintaining microbial populations and communities.
- Model, simulate, and predict the responses of microbes to each other and their environment by developing high-performance computational algorithms and infrastructure.
- Examine the dynamics of molecular machines in living cells.
- Measure gene expression and track the locations and interactions of proteins within living cells.
- Understand how individual cells function and interact within complex microbial communities to carry out complex processes.
- Integrate experiment, analysis, and theory in a recursive fashion to reveal intra- and intercellular networks.
- Enable development of microbial technologies and applications to solve DOE mission-specific problems in energy, environment, and health.

Project Purpose and Justification

High-throughput gene sequencing is being applied to determine the collective "genome" of microbial communities as a first step in understanding microbial community structural processes. Genome sequence alone, however, is insufficient to understand the functionality of microbes. A key to solving the formidable problem of understanding microbial genome function is through intensive experiment, analysis, and theory in a coupled and recursive manner. In a recent American Academy of Microbiology colloquium titled "Microbial Ecology and Genomics: A Crossroads of Opportunity," instrumentation development and new facilities were identified as critical needs in microbial science (see "AAM Recommends New Technologies," p. 7).

Facility IV will enable integration of the information and material outputs of the other GTL facilities to bring genomes to life. The facility will provide the tools, capabilities, and infrastructure needed to predict the functional behavior of whole cells and communities as integrated, dynamic systems. Facility IV will emphasize the concurrent and dynamic measurement of cellular proteins, molecular machines, intracellular metabolites, regulatory molecules, and gene transcripts to establish the state of cells within populations and communities and as a function of changes in physicochemical biological conditions.

Facility IV: Analysis and Modeling of Cellular Systems


Facility I
Reagents
Probes
Cell and community genomes
High-throughput Cultivation
Controlled experimental systems
Single-cell expression
Spatial, temporal analyses (imaging)
Data analysis
Data interpretation
Data archives
Modeling
Small-molecule analyses
New tools and infrastructure for unprecedented analyses of cell state
Cell and multicell models
Systems-level understanding of microbes and communities
Proteomics
Regulatory networks
Cultures
Communities
Machines for dynamic analysis
Cultures
Communities
Facility II
Facility III

Facility IV will combine advanced computational, analytical, and experimental capabilities for the integrated observation, measurement, and analysis of the spatial and temporal variations in the state of cellular systems—from individual microbial cells to complex communities and multicellular organisms. Facility IV will require the integrated information, materials, and capabilities provided by Facilities I-III as well as whole-genome sequence from the Joint Genome Institute. Facility I will provide critical affinity reagents and tags for single-cell measurements, while Facility II will furnish data on regulatory networks and high-throughput quantitative proteome and metabolome measurements. Facility III will produce information on key molecular machines as a basis for investigating their dynamics and functions in living cells. Facility IV, in turn, will provide insights into the function of molecular machines by defining their locations and dynamic behaviors within living cells. Facility IV also will be a source of microbial cultures and the knowledge about how to cultivate them.

Key Technologies Needed

- Cultivation and maintenance of microbes and microbial communities under controlled conditions, including the ability to interrogate the function of individual microbial cells in the context of a characterized physicochemical environment.
- Novel high-throughput cultivation approaches combined with single-cell analysis techniques, such as emerging microfluidic "lab-on-a-chip" devices, to grow and study currently uncultivable members of microbial communities.
- New multimodal capabilities for dynamic imaging of molecular machines and metabolites in individual cells and cell assemblies coupled with advances in computational resources for data acquisition, storage, and analysis to interpret and visualize the vast quantity of information obtained.
- Probes for the in situ measurement of extracellular metabolites and for defining the physicochemical environment in near real time.
- Analytical instrumentation and techniques for identifying and characterizing spatial and temporal variations in metabolites and signaling molecules, within and surrounding living microbial cells and cellular assemblies.
- Computational tools for efficiently collecting, analyzing, and integrating large data sets to elucidate gene function and to model and simulate regulatory and metabolic networks.
- New theory, algorithms, and implementation on high-performance computer architectures, such as those provided by the Ultrascale Simulation effort, to model and simulate cellular systems.
- Web- and grid-based technologies to enable a broad range of biological scientists to access the large data sets and computational resources needed for discovery-based biology.

Project Description

Facility IV will provide multiple capabilities to the scientific community. It will serve as a focal point for teams of scientists from academia, industry, and government to conduct integrated experiments on microbial processes and systems of interest and allow them to develop the detailed high-quality data sets under strictly controlled and characterized conditions required for elucidating regulatory and metabolic networks. Technologies for analyses at the single-cell level will require unique high-end instrumentation coupled to controlled cultivation systems with capabilities and tools for intensive data collection, integration, and analysis. These capabilities will be used in part or in their entirety by individuals or teams needing comprehensive analyses, at the single-cell or community level, of cellular systems of interest. This facility also will serve as fertile ground for training the next generation of scientists in systems biology and will attract students and faculty to work with multidisciplinary teams to reveal the functions of the microbial and microbial-community genomes.

Computational tools will be used to analyze large data sets and develop detailed models of cellular systems based upon these measurements. Computational methods ultimately will be used for large-scale simulations with these models to provide the ability to predict the behavior of cellular systems. These new capabilities will enable biologists to exploit emerging high-end computational tools for data analysis and the development of reliable predictive models. Facility IV will allow the fulfillment of GTL program goals for a systems-level understanding of microbes and microbial communities relevant to DOE missions in carbon cycling, metal and radionuclide bioremediation, and biomass conversion.

Impacts on Science and DOE Missions

Facility IV will be the capstone facility needed to provide the knowledge synthesis critical for bringing genomes to life. Understanding how individual subsystems of cells and individual cells within microbial communities function in concert to sense, respond to, and modify their environment represents a grand challenge for biology that must be addressed before scientists can successfully predict the behavior of microbes and take advantage of their functions.

In comparison to GTL Facilities I –III, which will provide new high-throughput production and analysis capabilities focused on defining and understanding the parts of microbial systems, Facility IV will focus on the dynamic systems-level study of living cells. It therefore will be highly data intensive, providing extensive linked data sets on the dynamic behavior of microbial cells and communities.

It also will be compute intensive, providing unprecedented data analysis, modeling, and simulation. These systems-level data sets will be made available to the scientific community and will be invaluable for identifying regulatory and metabolic networks in microbial systems and for advancing the annotation of microbial and community genome sequence. In addition, Facility IV will provide models and technologies for developing and evaluating such models and will provide user-facility type resources in the form of infrastructure needed to undertake such tasks.

Probabilities for Success

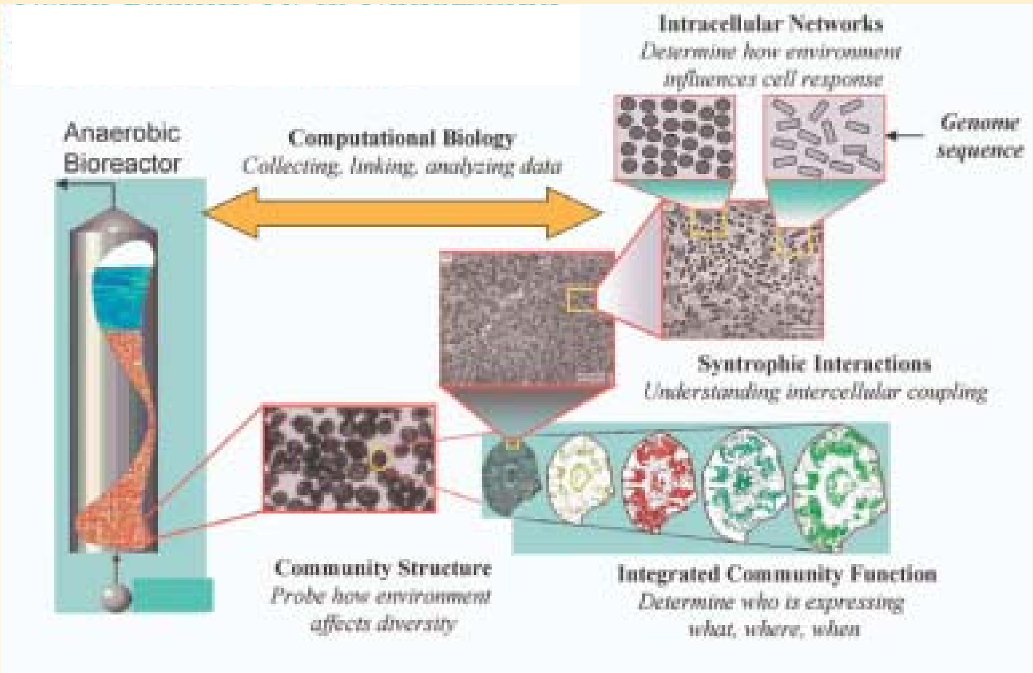
As in the genome projects, DOE can draw on multidisciplinary teams of biological, physical, computational, and other scientists and engineers from the national laboratories, academia, and industry to develop and deploy the resources for systematic functional genomic investigations of microbes and

A Possible Application of Knowledge Gained in GTL Facilities

Understanding Microbial Community Function is Essential for Using Bioreactors to Convert Waste to Clean Energy

Anaerobic bioreactors are used to treat wastewater and convert waste to "biogas" (mainly methane) that can be used as a clean energy source. Certain types of digesters contain self-aggregating biogranules composed of multiple spatially distinct microbial communities. The most widely used anaerobic treatment technology in the world, bioreactors have been the subject of extensive engineering studies. The biology, however, has largely been treated as a black box. The structure and function of microbial communities residing within the granules are believed to be important in maintaining an overall balance between production and consumption of metabolites and intermediates.

In Facility IV, the microbes, their molecular machines, and the complex interaction networks within and between cells in these communities would be revealed in detail. For example, the structure of biogranule communities, as well as integrated community function, would be established by determining which members are expressing particular genes or proteins at a given location and time. The integrated capabilities provided in this facility would allow scientists to "drill down" into these biogranule communities to probe intercellular coupling such as the transfer of metabolites that occurs between syntrophic microbial partners. Finally, at the level of the individual cell, scientists could determine how environment influences cell state and the dynamics of molecular machines within individual cells. A robust computational environment will be critical to Facility IV in collecting, linking, and analyzing experimental data and modeling and simulating complex systems such as the biogranule communities. The resulting information would have profound implications for controlling the efficiency and stability of methane-producing reactors and for greatly improving their design and operation.


Intracellular Networks
Determine how environment influences cell response
Genome sequence
Anaerobic Bioreactor
Computational Biology
Collecting, linking, analyzing data
Syntrophic Interactions
Understanding intercellular coupling
Community Structure
Probe how environment affects diversity
Integrated Community Function
Determine who is expressing what, where, when

microbial communities. DOE has an extensive and successful history of developing and applying new technologies to complex problems in the physical and chemical sciences. A tremendous opportunity now exists for applying these same talents to provide technological solutions to biology's most complex problems.

GTL Facility IV, more than the other three facilities, must take on and overcome major challenges associated with the lack of available technologies and instruments for measuring the dynamic state of living microbial cells. Facility IV will benefit from current and future R&D and pilot projects that will develop new technologies and instrumentation in a phased manner. These projects will provide new state-of-the-art capabilities by the time this facility comes on line. Facility IV also will include extensive capabilities for instrument and technology development that will be an essential part of this resource so that it can continue to measure the activities and characteristics of cellular systems at the single-cell level. In addition, it will involve development of new computational approaches for data storage, analysis, and use in complex models. For example, GTL-supported R&D and pilot projects under way include:

- Development of capillary analysis technologies to permit the monitoring of changes in protein expression in single cells using fluorescence, pushing the resolution by ten times over existing technology, is under way at the University of Washington. The goal is to build a "better microscope" for tracking gene expression in single cells following environmental changes.
- Electron tomography approaches are being developed at Lawrence Berkeley National Laboratory to image the inside of a microbial cell by freezing intact microbial cells in a way that preserves one layer of liquid water molecules above their membranes (permitting survival and viability). Electron microscopy images and computer reconstruction can then be used to derive 3D images of internal cell constituents.
- A pilot Microbial Cell Dynamics Laboratory is under development at Pacific Northwest National Laboratory to provide flexible experimental systems to control and manipulate microbial growth conditions and to make multiplexed measurements of cellular activities and responses to changes in environmental conditions.

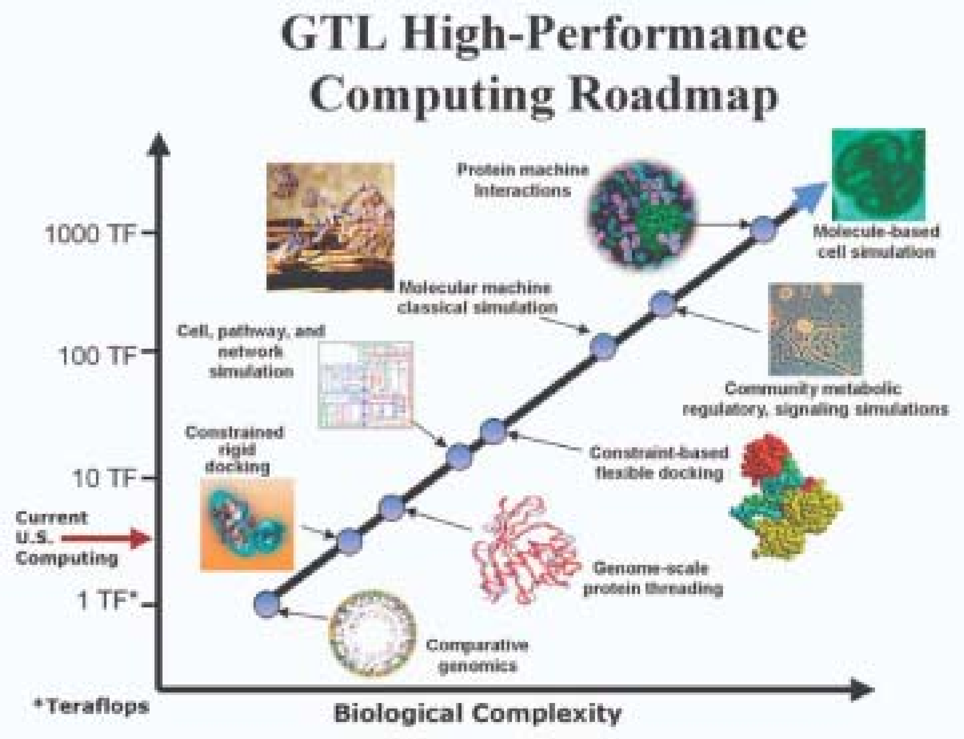
Cross-Cutting Computing Infrastructure for GTL Facilities and Projects

Strategic Intent

Biology is poised on the threshold of a great transformation that will create a new life-science approach in which large data sets and advanced computing will be combined into predictive simulations to guide and interpret experimental studies. This new biology will allow a level of understanding that will enable biological solutions to many of the world's pressing challenges including energy security, control of atmospheric carbon, and environmental cleanup.

Such progress depends on the emergence of a new mathematical, quantitative, predictive, and ultimately systems-level paradigm for the life sciences. This new paradigm is one in which biologists represent their most fundamental knowledge of complex biological systems as mathematically based computer models. These models will be used to capture and represent data, predict behavior, and generate hypotheses that can be tested by gathering more data (see "Biology Paradigm," p. 34). Biologists need to be provided with the means to move data and knowledge back and

Important breakthroughs in GTL modeling can be made using the next generation of high-performance computing platforms with 40- to 100-TF capability, including docking and simulation of protein-protein interactions using classical energy functions. More complex models and simulations of machines and cellular systems would benefit greatly from fundamental research in mathematics and from related algorithms that could dramatically improve calculation efficiencies compared to current estimates. Constraints (bounds and guides) provided by data such as observed machine geometries will make calculations of more complex systems tractable.



The so-called First Principles Molecular Dynamics (FPMD) methods, the current state-of-the-art in biophysical modeling, simulate the motions of atoms in biochemical systems using a quantum mechanical description of atomic interactions. Highly optimized for DOE's current teraflop-speed computers, FPMD applications are capable of simulating up to a few hundred atoms for a few picoseconds. These methods, however, also constitute a nearly exact simulation of nature, and, even within these computational limitations, FPMD is becoming an important tool for studying fundamental biochemical processes. The results for small biochemical systems currently being simulated on teraflop-scale computers provide tantalizing glimpses of the value of longer-time and larger-system simulations that will be made possible with faster computers.

forth between experiment and computing on an everyday basis, making large-scale computing an integral part of their daily lives. This will be necessary to study even the simplest microbes at a level of detail sufficient to predict their behavior. To ask next-generation questions and do next-generation experiments, computing must guide the questions and interpretation at every step.

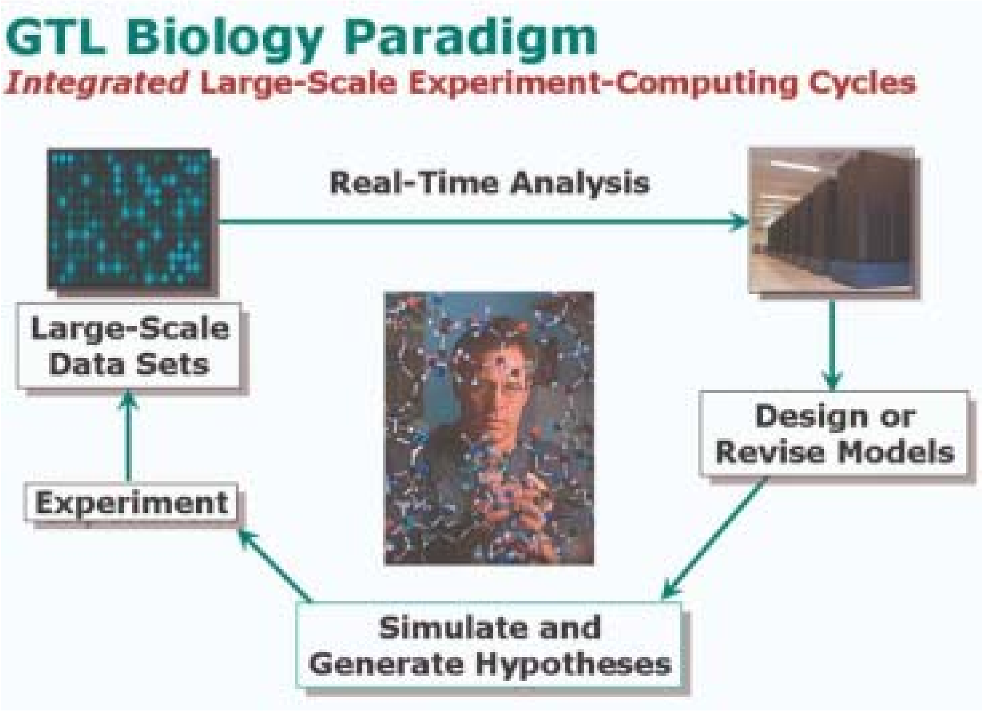
Adopting this new paradigm to meet the requirements of GTL facilities is the result of the following specific drivers.

- **Data Analysis.** Production facilities in Genomes to Life will generate vast amounts of diverse and complex data that must be analyzed, integrated, and interpreted. The complexity of data-generation modalities is much higher than in the genome era, and the amounts of data will be much larger than for sequencing the human genome. Achieving the necessary data-analysis throughput will require significant advances in software and hardware infrastructure.
- **Complexity.** The complexity of systems in even the simplest microbes makes manual analysis and annotation methods simply inadequate. This mind-boggling complexity needs to be captured in the computer and denote biology in mathematical ways that parameterize system complexity. The need to capture, represent, and model complex biology via computer language is fundamental to progress in Genomes to Life facilities and projects.
- **Prediction, Quantitation, and Simulation.** Quantitating, predicting, and simulating behavior are necessary to understand biological systems and develop hypotheses for further testing. Genomes to Life goals will require simulation of heterogeneous biological systems over long time scales and include molecular complexes, pathways, networks, and, eventually, communities of organisms. Quantitation is needed to test our understanding and representation of systems.
- **Principles and Concepts.** To understand the astounding complexity of biological systems, general systems principles need to be extracted and developed from systems data, modeling, and simulation. Mathematical representations of complex biological systems are fundamental to the conceptual breakthroughs anticipated in Genomes to Life.

Purpose and Justification

Computing capabilities for analyzing, modeling, and simulating the dynamic behavior of complex biological systems are an essential complement to GTL experimental data-collection activities that provide fundamental observations and data. In addition to providing a totally new capability for understanding the basis of life at its most fundamental level, the ability to compute and simulate large macromolecular machines and systems is central to many DOE missions, including bioremediation, climate, and energy security.

In bioremediation and climate, interactions of microbes with the environment and with each other also involve large complexes of macromolecules and metabolic systems that act on pollutants and sense environmental and community conditions. Data analysis followed by modeling and simulation can provide fundamental understanding as well as quantitative parameters important for models of larger earth systems such as climate.



In energy security, the potential for biological energy production is very great. Our understanding and ability to engineer complex systems that synthesize useful compounds would be enhanced by a fundamental grasp of the proteins and metabolic processes involved and the ability to predict the behavior and dynamics of these molecules and systems.

Software, Data, Biocomputing Centers

The Genomes to Life facilities plan and workshop reports place considerable emphasis on developing methods for a large community of biologists to analyze petascale biological data sets and develop models and simulations related to complex biological phenomena. They stress the need for centrally managed approaches to software and hardware infrastructure to accomplish these objectives. A centralized approach to coordination and planning in computing will guide data management, large-scale development of analysis tools, implementation, and support of analysis on specialized hardware environments, including massively parallel computers and distributed grid systems.

Components of the software-development and data-management infrastructure, such as GTL databases, will be designed around and co-located with major facilities that generate biological data. Experience in many large-scale genome-era projects, such as the Joint Genome Institute, has shown that physical co-location is immensely useful and absolutely necessary for effective communication and requirement definition among biologists and computational scientists. While the overall goal is to create a seamless and effectively centralized capability to deal with data, software development, and high-performance computing, key development teams will be based primarily at GTL experimental data facilities and coordinate their activities across the GTL enterprise.

The need is significant for large-scale compute tools and resources that will become a shared resource for the GTL community. These “centers” will, in many cases, be distributed within several GTL facilities or other sites. Three different types of components are required: (1) coordinated centers for applications software and tools development, (2) data-management and -integration resource centers, and (3) biocomputing centers that support community access, analysis, modeling, and simulation using a specialized computing hardware.

Several key points about this strategy are the following: (1) Domain experts who team to develop applications do not have to be co-located at sites where substantial computer hardware exists. (2) Centralizing management of applications ensures that they are supported, shared, and documented. (3) Analysis, modeling, and simulation applications can find cycles from multiple sites and are not as vulnerable to individual machine failures. (4) Although many users will be experts and understand the machine requirements of their codes, most biological users will not. To facilitate wide usage of GTL infrastructure in computing, very simple user environments must be created that “know” where an application should run (on what type of platform) and where to get the necessary data without the user having to specify these details. (5) By sharing compute hardware resources across the GTL enterprise and among GTL facilities, administrative processes can more effectively use a variety of machine environments, from large clusters to massively parallel processing (called MPP) machines, as the demand for processing dictates. Applications can be matched to the most appropriate environments.

Centers for the Development of Analysis, Modeling, and Simulation Tools

GTL data generation and computing advances will provide scientists with access to comprehensive information and the tools to incorporate it into models to probe the processes and phenomena of living systems, test hypotheses and ideas, and inspire and inform new types of experimental inquiry. Tool centers develop, maintain, and support analysis and modeling-code repositories. They collect, develop, curate, and implement analysis, modeling, and simulation tools related to GTL tasks and make them available to biology users at GTL centers and in the community. Tool centers would provide a tool repository accessible to investigators or directly by machines in the national grid. For example, users could go to an access point and specify that a tool from a particular repository be used for a task. By coupling activities at GTL facilities and other sites, tool centers would focus on several types of analysis applications:

- Bioinformatics Tools
  - Microbial-community sequence analysis and annotation

  - Proteome and expression-data analysis
- Biophysics Tools
  - Protein dynamics and protein chemistry
  - Protein docking, protein machine modeling and simulation
- Biosystems Tools
  - Metabolic modeling
  - Cell and regulatory-network modeling

Data Centers

Key to GTL's success is genome-scale collection, analysis, dissemination, verification, and modeling of data. Just as with the Human Genome Project and community production of DNA sequence, a key to GTL's success will be the generation of genome-scale data and the data-management and -analysis capabilities needed to interpret the biological "outputs" of a genome. Centrally coordinated data centers, most often located at closely related GTL facilities, would accumulate and integrate data from GTL facilities and distributed projects and organize it for use by the community and GTL modelers. Mirrors of these databases would be supported at biocomputing centers, where the data would be available for incorporation into analysis processes.

Several types of major data resources likely to be needed should be designed in a coordinated way:

- Expression and proteomics databases
- Protein-function and protein-chemistry databases
- Protein-machine, protein-complex, and dynamics database
- Metabolic-pathway and pathway-model database
- Regulatory-network and cell-modeling database
- Microbial-community sequences and annotation database

Biocomputing Centers

The path to understanding the function and dynamic behavior of large molecular systems involves computing, modeling, and simulation of these systems based on structure data, informational parameters, and the use of physically based principles and methods. Biocomputing centers will pool specialized high-performance resources and distributed cluster hardware to provide user access to environments that facilitate large-scale analysis, modeling, and simulation processes. These centers will share a relatively uniform suite of applications (obtained from tool development in GTL facilities and projects) and also mirror databases needed for various analyses or simulations.

While single-molecule simulations currently can be achieved in about a microsecond, the dynamics of protein-protein interactions and simulations of even larger complexes of macromolecules will require much more computing capability. The scale of such simulations poses a significant challenge, requiring capabilities from 50 teraflops to petaflops and beyond (see graphic, p. 31). With a focus on achieving this infrastructure in the next 5 to 10 years, tremendous breakthroughs can be obtained in our understanding of the most fundamentally important macromolecular machines. A similar and somewhat parallel set of requirements will apply to other GTL areas such as network, pathway, and cell modeling.

The infrastructure needed to support computing, modeling, and simulation in GTL facilities and projects will strongly leverage the continuing development of high-performance computing capability within the DOE Advanced Scientific and Computing Research program.

Genomes to Life Program and Facilities Planning Workshops

A series of program-planning workshops has been held to help plan and coordinate Genomes to Life. Meeting reports are placed on the Web as soon as they become available. To learn more about the program, please see the Web site (DOEGenomesToLife.org).

Web site for workshop reports: DOEGenomesToLife.org/pubs.html

2000

October 29–November 1	Genomes to Life Roadmap Planning, San Diego

2001

January 25–27	Genomes to Life Roadmap Planning, Germantown, Md.
June 23	Role of Biotechnology in Mitigating Greenhouse Gas Concentrations, Arlington, Va.
August 7–8	Computational Biology, Germantown, Md.
September 6–7	Computational and Systems Biology, Washington, D.C.
September 9–10	Science Mission Payoffs, Washington, D.C.
October 24–25	Energy and Climate Mission Payoffs, Chicago
December 10–11	Technology Assessment for Mass Spectrometry, Washington, D.C.

2002

January 22–23	Computational Infrastructure, Gaithersburg, Md.
March 6–7	Computer Science, Gaithersburg, Md.
March 18–19	Mathematics, Gaithersburg, Md.
April 16–18	Imaging, Charlotte, N.C.
April 16–19	Computing Strategies, Oak Ridge, Tenn.
June 19–20	Facilities Planning, San Francisco
August 16–17	Facilities Planning, Chicago
October 14–15	Facilities Planning, Gaithersburg, Md.

Entities and Institutions Represented at GTL Workshops and Meetings

Affymetrix • Ames Laboratory • Argonne National Laboratory • Bell Labs • Boston University • Brookhaven National Laboratory • California Institute of Technology • Carnegie Mellon University • Celera Genomics • Columbia University • Cornell University • Dana-Farber Cancer Institute • Duke University • Duke University School of Medicine • DuPont • East Carolina University • Energy Sciences Network (Esnet) • Food and Drug Administration • Genentech Inc. • General Electric • geneticXchange • Harvard Medical School • Harvard University • Hebrew University • InPharmix Inc. • Institute for Systems Biology • IBM • Jefferson Lab • Johns Hopkins School of Medicine • Johns Hopkins University • Joint Genome Institute • Joint Institute for Computational Science • Keck Graduate Institute • Keio University • Lawrence Berkeley National Laboratory • Lawrence Livermore National Laboratory • Los Alamos National Laboratory • Marshfield Medical Research Foundation • Massachusetts Institute of Technology • Medical University of South Carolina • Merck Research Laboratories • Molecular Sciences Institute • Monsanto Company • Montana State University at Bozeman • Monterey Bay Aquarium Research Institute • National Academy of Sciences • National Cancer Institute • National Center for Biotechnology Information • National Center for Genome Research • National Center for Supercomputing Applications • National Energy Research Scientific Computing Center • National Human Genome Research Institute • National Institute of General Medical Sciences • National Institutes of Health • National Renewable Energy Laboratory • National Research Council • National Science Foundation • National Water Research Institute • Natural Resources Defense Council • New England Complex Systems Institute • New York University • North Carolina Supercomputing Center • Novation Biosciences • Oak Ridge Institute for Science and Education • Oak Ridge National Laboratory • Office of Management and Budget • Ohio State University • Pacific Northwest National Laboratory • Pittsburgh Supercomputing Center • Princeton University • Rockefeller University • Sandia National Laboratories • Sanger Centre • Scripps Institution of Oceanography • Scripps Research Institute • Southwest Parallel Software • SRI International • Stanford School of Medicine • Stanford University • Test Measurement Systems Inc. • Texas Tech University • The Institute for Genomic Research • The Packson Laboratory • United State Department of Agriculture • University of California, Berkeley • University of California, Irvine • University of California, Los Angeles • University of California, San Diego • University of California, San Francisco • University of California, Santa Barbara • University of Colorado • University of Connecticut Health Center • University of Florida • University of Illinois • University of Illinois at Urbana-Champaign • University of Iowa • University of Maryland Biotechnology Institute • University of Massachusetts, Amherst • University of Miami • University of Michigan • University of Pennsylvania • University of Pittsburgh • University of Southern California • University of Texas • University of Utah • University of Washington • University of Wisconsin • Vanderbilt University • Vertex Pharmaceuticals Inc. • Weyerhauser Company • Whitehead Institute for Genome Research

U.S. Department of Energy Office of Science

Marvin Frazier
Office of Biological and Environmental Research (SC-72)
301/903-5468, Fax: 301/903-8521
marvin.frazier@science.doe.gov

Gary Johnson
Office of Advanced Scientific Computing Research (SC-30)
301/903-5800, Fax: 301/903-7774
garyj@er.doe.gov